Tarea 4: Bootstrapping método del percentil

Instituto Tecnológico Autónomo de México
Maestría en Ciencias en Computación
Karen L. Poblete 116452
Estadística Computacional

26 de octubre de 2015

1. Introducción

Bootstrapping fue introducido por vez primera en 1979 por Efron [3]. Desde ese entonces numerosas publicaciones han sido escritas al respecto, entre las que se pueden citar [4], [5], [6], y [7], entre otras. Los métodos conocidos como bootstrap son una clase de métodos de Monte Carlo basado en simulaciones no paramétricos, es decir, no se asume alguna distribución en particular para el conjunto de muestras. Lo que se hace es estimar la distribución de una población mediante el remuestreo.

Como se ha mencionado con anterioridad, bootstrapping es un método de remuestreo utilizado para inferir la distribución muestral de un estadístico de un conjunto de variables aleatorias pertenecientes a una distribución específica. El método de Bootstrapping se utiliza comúnmente para estimar características de interés de muestras de una distribución particular como podrían ser la media, sesgo y la varianza, además de hacer comparaciones entre estadísticos de diferentes poblaciones. También es utilizado para crear intervalos de confianza de hipótesis de parametros de interés.

Bootstrap trata a una muestra observada como una población finita, y las muestras aleatorias son generadas para estimar las características de la población y posteriormente hacer inferencias sobre la población muestreada. Estos métodos son frecuentemente utilizados cuando la distribución de la población objetivo no está especificada, es decir, sólo se tiene información de la muestra. Existen dos tipos de bootstrap, el paramétrico y el no paramétrico, cuando se utiliza el paramétrico involucra remuestreo a partir de una distribución de probabilidad completamente específicada [9].

El método básico de bootstrap se puede definir a partir de tener una muestra aleatoria $x = (x_1, x_2, ..., x_n)$ observada a partir de una $F(x)$. Si X^* es seleccionada aleatoriamente de x, entonces,

$$P(X^* = x_i) = \frac{1}{n},\ i = 1, 2, .., n$$

Con remuestreo se genera una muestra $X_1^*, X_2^*, ..., X_n^*$ mediante muestreo con reemplazo de x. Las variables aleatorias X_i^* son i.i.d, y uniformemente distribuidas en $x_1, x_2, ..., x_n$. Ahora bien, $F_n(x)$ es la función de distribución empírica (ecdf) y es un estimador de $F(x)$. Bootstrap propone hacer un análisis exahustivo de la distribución muestral del estadístico de interés por medio de tomar submuestras de la muestra.Como dicha actividad tomaría mucho tiempo, en casos reales, se procede a tomar B submuestras y por medio de ellas llegar a una buena aproximación del estadístico deseado. Como es de suponerse existe un error inherente al método de bootstrap.

2. Estimación bootstrap del error estándar y sesgo

La estimación bootstrap del error estándar de un estimador $\hat{\theta}$ es la desviación estándar muestral de las réplicas $\hat{\theta^{(1)}}, \hat{\theta^{(2)}}, .., \hat{\theta^{(B)}}$ definida por:

$$\hat{se}(\hat{\theta^*}) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta^{(b)}} - \bar{\hat{\theta}}^*)^2}$$

Donde $\bar{\hat{\theta}}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta^{(b)}}$. De acuerdo a [4], el número de rélicas necesarias para tener buenas estimaciones del error estándar no es muy grande, basta con que $B \approx 50$. Para intervalos de confianza de recomienda $B > 200$. El objetivo es encontrar estimadores con el menor sesgo posible al valor real del parámetro de interés.

Si $\hat{\theta}$ es un estimador insesgado de θ, $E[\hat{\theta}] = \theta$. Su sesgo viene dado por:

$$sesgo(\hat{\theta}) = E[\hat{\theta} - \theta] = E[\hat{\theta}] - \theta$$

Cada estadístico es un estimador insesgado de su propio valor esperado, en particular, la media muestral de una v.a. es un estimador insesgado de la media de la distribución. La estimación bootstrap del sesgo usa las réplicas de $\hat{\theta}$ para estimar la distribución muestral de $\hat{\theta}$. Para la población finita $x = (x_1, x_2, ..., x_n)$, el parámetro es $\hat{\theta}(x)$ y hay B estimadores i.i.d $\hat{\theta^{(b)}}$. La media muestral de las réplicas $\hat{\theta}^{(b)}$ es insesgado por su valor esperado $E[\hat{\theta}^*]$, así que el estimador bootstrap para el sesgo es:

$$\widehat{sesgo(\widehat{\theta})} = \bar{\widehat{\theta}}^* - \widehat{\theta}$$

Donde $\bar{\widehat{\theta}}^* = \frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}^{(b)}$ y $\widehat{\theta} = \widehat{\theta}(x)$ es la estimación calculada a partir de la muestra original observada.

3. Intervalos de confianza con Bootstrap

Bootstrap también es utilizado para construir pruebas de hipótesis. Se utiliza cuando no se puede asumir por completo que se está utilizando una distribución paramétrica en el remuestreo. Muchas veces dicha afirmación es puesta en duda y por tal motivo se pueden construir intervalos de confianza. También se utiliza en casos donde su cálculo es dificil de realizar debido a falta de recursos computacionales o es muy complejo matematicamente. Existen varios métodos para construir dichos intervalos de confianza.

3.1. La normal estandar

Supongamos que $\hat{\theta}$ es un estimador de θ con error estándar $se(\hat{\theta})$. Si $\hat{\theta}$ es la media muestral y ésta es grande, entonces el Teorema Central del Límite implica que [1]:

$$Z = \frac{\hat{\theta} - E[\hat{\theta}]}{se(\hat{\theta})} \sim N(0,1)$$

Si $\hat{\theta}$ es un estimador insesgado, entonces un intervalo de confianza del $100(1-\alpha)\,\%$ es:

$$\hat{\theta} \pm z_{\frac{\alpha}{2}} se(\hat{\theta})$$

Donde $z_{\frac{\alpha}{2}} = \Phi^{-1}(1-\frac{\alpha}{2})$. Para poder utilizar este método es necesario cumplir con los siguientes supuestos: (i) la distribución de $\hat{\theta}$ es normal o bien, $\hat{\theta}$ es la media muestral de una muestra muy grande. (ii) $\hat{\theta}$ es un estimador insesgado, es decir, su sesgo es nulo por sr su esperanza igual a θ. (iii) $se(\hat{\theta})$ es desconocido, pero con bootstrap esto corresponde a la desviación estándar de las réplicas.

3.2. Método sencillo

Transforma la distribución de las réplicas mediante la substracción del estadístico observado. Se usan los cuantiles de la muestra transformada para determinar los límites de confianza.

$$(2\hat{\theta} - \hat{\theta}_{1-\frac{\alpha}{2}}, 2\hat{\theta} - \hat{\theta}_{\frac{\alpha}{2}})$$

3.3. Método de percentiles

Usa la distribución empírica de las réplicas bootstrap como la distribución de referencia. Los cuantiles de la distribución empírica son estimadores de os cuantiles de la distribución de $\hat{\theta}$, de manera que pueden corresponder más a la distribución verdadera cuando $\hat{\theta}$ es no normal. Sean $\hat{\theta}^{(1)}$, $\hat{\theta}^{(2)}$,...,$\hat{\theta}^{(B)}$ las réplicas de $\hat{\theta}$. A partir de su función de distribución estimada se calcula el cuantil $\frac{\alpha}{2}$, $\hat{\theta}_{\frac{\alpha}{2}}$, y el cuantil $1-\frac{\alpha}{2}$, $\hat{\theta}_{1-\frac{\alpha}{2}}$. En [4] se muestra que este tipo de intervalos tiene ventajas teóricas y además de una mejor representación.

Para hacer pruebas de hipótesis con bootstrap se debe basar el valor del p-value en el intervalo de confianza se generó. Específicamente, se debe considerar una hipótesis nula expresada en términos de un parámetro cuya estimación se puede hacer por medio de bootstrap. Si $(1-\alpha)100\,\%$ intervalo de confianza de bootstrap para el parámetro no cubre el valor obtenido, entonces se rechaza la hipótesis nula con un valor p-value no mayor a α [1].

El método de percentil puede traer error, ya que se está comparando al valor muestral del estimador de interés. Para hacerlo más certero, debería ser pivotal, es decir, conocer algunos parámetros de la distribución del parámetro que se quiere aproximar con el estimador de interés. De ahí surge el uso de la creación de intervalos de confianza utilizando la distribución t [1].

4. Código ejemplo de método de percentiles R

```
data(patch)
theta.boot<-function(data,ind){
y<-data[ind,1]
z<-data[ind,2]
mean(y)/mean(z)}
y<-patch$y
z<-patch$z
data<-cbind(y,z)
boot.out<-boot(data,statistic=theta.boot,R=5000)
print(boot.out)
print(boot.ci(boot.out,type=c("basic","norm","perc")))
```

5. Conclusiones

Al utilizar métodos de remuestreo como bootstrap siempre existe un error inherete de aproximación a la distribución muestral de un parámetro en específico de una función cuya distribución sea $F(x)$. Debido a esto es necesario comprobar que dicho estimador obtenido por bootstrap se aproxima al valor que desea estimar. Un método para realizar esto, es por medio de intervalos de confianza, donde esperamos que el valor obtenido caiga en por lo menos $(1-\alpha)100\,\%$ de la distribución muestral. Así podemos afirmar que tenemos una aproximación aceptable. El método de percentiles es el más sencillo para realizar ésta clase de pruebas de hipótesis ya que se basa en la función de densidad muestral del estimador y no se requiere hacer cálculos complecados. Además el resultado puede ser manegado con facilidad, pero podría ocultar sesgo ya que compara los valores estimados contra la media de la distribución muestral del estimador.

Referencias

[1] Givens, G., & Hoeting, J. *Computational statistics*. John Wiley & Sons, 2012.

[2] Mauricio García Tec. *Apuntes de Estadística Computacional*. ITAM, 2015.

[3] Efron, B. *Bootstrap methods: another look at the jackknife*. The annals of Statistics, 1979, p. 1-26.

[4] Efron, B., & Tibshirani, R. *An introduction to the bootstrap*. Monographs on statistics and applied probability, 1993, vol. 57.

[5] Efron, B. *Nonparametric standard errors and confidence intervals*. The Canadian Journal of Statistics/La Revue Canadienne de Statistique, 1981, p. 139-158.

[6] Suzuki, R., & Shimodaira, H. *Hierarchical clustering with P-values via multiscale bootstrap resampling*. R package, 2013.

[7] Deng, N., et al. *Using the bootstrap to establish statistical significance for relative validity comparisons among patient-reported outcome measures*. Health Qual Life Outcomes, 2013, vol. 11, p. 89.

[8] Tong, L., et al. *Quantifying uncertainty of emission estimates in National Greenhouse Gas Inventories using bootstrap confidence intervals*. Atmospheric environment, 2012, vol. 56, p. 80-87.

[9] Wasserman, L. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013.